Sentence Repetition as a Clinical Marker for Mandarin-Speaking Preschoolers with Developmental Language Disorder

Danyang Wang[1], Li Zheng[2], Yuanyuan Lin[3,4], Yiwen Zhang[3,4], Li Sheng[5]

[1]Department of Communication Sciences and Disorders, University of Delaware, Newark, Delaware, USA; [2]Nanjing Normal University, Nanjing, Jiangsu, China; [3]Developmental and Behavioral Pediatrics Department, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [4]Ministry of Education-Shanghai Key Laboratory of Children's Environmental Health, Xinhua Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; [5]Research Centre for Language, Cognition, and Neuroscience & Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China.

Address correspondence to Li Sheng, Research Centre for Language, Cognition, and Neuroscience & Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China. Email: dr-li.sheng@polyu.edu.hk. Tel: 852-27667446, Fax: 852-23340185.

The authors have no relevant conflicts of interest to declare.

This project was funded by Humanities and Social Sciences projects of the Chinese Ministry of Education (17YJAZH132) awarded to Li Zheng (PI) and Li Sheng (co-PI), and a Pudong One Hundred Award to Li Sheng.

Abstract

Purpose: Sentence repetition (SR) is believed to be a clinical marker for Developmental Language Disorder (DLD) across many languages. This study explored the potential of a self-designed Mandarin SR task (MSRT) to reflect Mandarin-speaking preschoolers' language ability and to differentiate children with and without DLD in this population. Furthermore, we aimed to compare five scoring systems for evaluating children's MSRT performance. **Method**: In study 1, the MSRT was administered to 59 typically-developing (TD) children aged 3;6 (years; months) to 6;5 in China. The task was examined regarding its ability to correlate with language indices derived from children's narrative samples. In study 2, both a TD and a DLD group were recruited to investigate the task's sensitivity, specificity, and likelihood ratios to distinguish between children with and without DLD. **Results:** Study 1 showed that, using four of the five scoring methods, TD children's performance on the MSRT significantly correlated with all the language measures derived from narratives. Study 2 showed that the MSRT was able to differentiate children with and without DLD. **Conclusion:** The MSRT is a promising tool to reflect language abilities and identify DLD in Mandarin-speaking preschoolers. Based on the current evidence, we recommend that researchers and clinicians select the number of errors in syllable method or the binary method when scoring responses to meet their specific needs.

Keywords: sentence repetition, clinical marker, Mandarin, developmental language disorder

Introduction

Children with Developmental Language Disorder (DLD, also known as Specific Language Impairment), demonstrate significant deficits in talking and/or understanding language, and these deficits are not attributed to any physical or neurological conditions (Bishop et al., 2017). DLD negatively affects children's everyday life and school achievement, including social interaction, literacy, and mathematical thinking (Knox & Conti-Ramsden, 2003; McArthur et al., 2000). Although approximately 7% of children are affected by DLD (Norbury et al., 2016; Tomblin et al., 1997), this disorder is notoriously under-detected in real life (e.g., Jessup et al., 2008; Tomblin et al., 1997). There is a pressing need to develop effective screening methods so that children who are at risk of having DLD can receive further diagnostic assessments which allow them to receive timely support from speech-language pathologists and teachers. A good screening task needs to have high sensitivity to not miss any potential cases of DLD; it also needs to be timesaving so that it can be administered at scale. Sentence repetition (SR, also termed sentence recall, sentence imitation, recalling sentences) tasks are good candidates for this purpose, given its utility in differentiating children with and without DLD across languages (e.g., Conti-Ramsden et al., 2001; Redmond, 2005; Stokes et al., 2006) and its quick administration and scoring. This paper aims to develop and validate a Mandarin SR task as a measure of language abilities and evaluate its classification accuracy in differentiating Mandarin-speaking preschool children with and without DLD.

Nature of the SR task

Various SR tasks have been developed in a myriad of languages. An SR task involves having speakers listen to auditorily-presented sentences one at a time and repeat each sentence verbatim immediately after presentation. This task is widely recognized as a useful measure of individual differences in speakers' language ability (Polišenská et al., 2015). Baddeley (2000) suggested that speakers rely on their long-term semantic and grammatical knowledge to enable the binding of words into larger sentence-level chunks when performing the SR task. In support of this, Klem et al. (2015) conducted a study with 216 children and found significant correlations between children's SR performance and their knowledge of vocabulary and grammar. The authors suggested that the SR task is a complex language task that reflects the integrity of language processing systems at multiple levels, including lexical and grammatical skills, as well as speech perception and speech production. Similarly, Polišenská et al. (2015) identified the involvement of lexical knowledge and morphosyntax in the successful repetition of sentences and claimed that the SR task reflects speakers' general language ability. Although phonological memory is recruited when performing the SR task (Alloway & Gathercole, 2005), Archibald and Joanisse (2009) found that their SR task was more sensitive to deficits in linguistic rather than memory abilities.

Not only is the SR task sensitive to individual differences in spoken language ability, it is also a good candidate for identifying children with DLD, as it heavily recruits skills that are known to be weak in this population, such as vocabulary, grammar, and phonological memory (e.g., Bishop et al., 2016; Leonard, 2014; Trauner et al., 1995). Children with DLD start expressing meaning with words 11 months later than their typical peers do (Trauner et al., 1995). Throughout preschool years, they continue to demonstrate deficits in receptive vocabulary (Bishop, 1997; Clarke & Leonard, 1996), expressive vocabulary (Leonard et al., 1999; Thal et al., 1999; Watkins, 1995), and novel word learning (Kan & Windsor, 2010). Early school-age children with DLD also demonstrate word retrieval and semantic processing deficits (Sheng, 2014). Grammatical (both morphology and syntax) difficulty is a hallmark deficit in individuals with DLD cross-linguistically (Leonard, 2014). English-speaking children with DLD have difficulties using past tense (e.g., walked) and plural inflections (e.g., ducks) (Joanisse & Seidenbert, 1998), following appropriate word orders (Hansson & Nettelbladt, 1995), producing wh- questions (Van der Lely & Battell, 2003), and using adjuncts (Johnston & Kamhi, 1984). Across different languages, children with DLD show grammatical difficulties that are specific to the ambient language, such as inflectional morphology in Hungarian (Leonard et al., 2009) and aspect markers in Chinese (Fletcher, 2005; Hao et al., 2018). An SR task that capitalizes on these areas of known deficits could therefore act as an effective identification tool of DLD.

Use of Sentence Repetition for Identifying Individuals with DLD

Multiple studies have shown that SR is a clinical marker of DLD, by testing children with and without DLD and investigating the task's classification accuracy values, including sensitivity and specificity. Sensitivity refers to the test's ability to accurately capture individuals with DLD, and specificity reflects the test's ability to accurately identify TD individuals. For diagnostic tasks, Plante and Vance (1994) proposed a guideline which considers sensitivity and specificity values below .80 as unacceptable, values of .80-.89 as acceptable, and values at or over .90 as good.

Different SR scoring methods were explored in the literature and were shown to affect the classification accuracy of SR tasks. Commonly used scoring systems include (1) correct/incorrect scoring method (binary method; Newcomer & Hammill, 2019; Rispens, 2004), which gives a score of one to completely accurate repetitions and a score of zero to responses with any deviations; (2) scoring that considers errors (error scoring method), in which the score (0-2 in Redmond, 2005; 0-3 in the *Clinical Evaluation of Language Functions*, CELF, Wiig et al., 2013) is based on the number of errors in the response; (3) scoring that considers the specific grammatical structures in the sentences (known as the core element scoring or grammatical scoring, Komeili & Marshall, 2013), which gives a score of one to the response that contains the target grammatical structure and a score of zero to a response without the target structure; (4) scoring that calculates the percent of correct syllables (correct syllable/total syllables; Stokes et al., 2006).

Research evidence has supported the use of SR tasks as clinical markers for DLD in English. Conti-Ramsden et al. (2001) compared four potential psycholinguistic markers for DLD in English-speaking children: an SR task (CELF-R-Recalling Sentences subtest, Semel et al., 1994), a nonword repetition task (in which children repeat nonsense words of varied lengths), a third person singular task, and a past tense task. Compared to the other three markers, the SR task demonstrated the highest diagnostic accuracy (using the error scoring), with a sensitivity of 90% and a specificity of 85% (using a cutoff score of -1SD). Using the error method, another English SR task developed by Redmond (2005) also demonstrated good utility (sensitivity of 94% and specificity of 88%) in differentiating children with and without DLD.

SR tasks were shown to be effective clinical markers for DLD in other languages as well. Leclercq et al. (2014) compared group performance in children with and without DLD on a French SR task under seven different scoring methods[1]. Children with DLD performed significantly lower than TD children on the French SR task, regardless of the scoring method. When using the correct/incorrect scoring method, the discriminant function analysis revealed the highest levels of sensitivity (97%) and specificity (88%). Furthermore, Armon-Lotem and Meir

[1] The seven scoring methods included the binary scoring; grammatical scoring; scoring that considers three core semantic ideas in each sentence; as well as scorings that calculate the number of correct words; number of correct morphemes; number of function words; and number of content words.

(2016) established the effectiveness of their SR tasks in distinguishing monolingual children with and without DLD in both Russian (sensitivity of 86%, specificity of 90%) and Hebrew (sensitivity of 100%, specificity of 87%), although the scoring method was not specified.

Only a few studies have investigated the utility of SR as a clinical marker for DLD in Asian languages, including Korean (Hwang, 2012), Vietnamese (Vân Hoàng et al., 2014; Pham & Ebert, 2020), and Cantonese (Stokes et al., 2006). Pham and Ebert (2020) tested 104 Vietnamese-speaking five- and six-year-old children, including ten children with DLD. The authors explored three scoring methods: binary, error, and grammatical scoring. When using the error scoring method, their SR task achieved a sensitivity of 90% (CI[2]: 0.71 to 1.09[3]) and a specificity of 71% (CI: 0.43 to 0.99). The binary method yielded slightly higher sensitivity (100%) but lower specificity (57%, CI: 0.26 to 0.88) than the error method. When using the grammatical scoring method, the sensitivity was 80% (CI: 0.55 to 1.05) and the specificity was 71% (CI: 0.43 to 0.99).

Stokes et al. (2006) examined a Cantonese SR task with a DLD group (N=14), an age-matched TD group (N=15) and a younger language-matched TD group (N=15). Four different scoring methods were explored: binary, error, grammatical and percent of correct syllable scoring. The age-matched TD group performed significantly higher than the language-matched TD group and the DLD group, whereas the latter two groups did not differ from each other, regardless of the scoring method. Classification accuracy was evaluated using the error method and the percent correct syllable method. The error method resulted in higher classification values (sensitivity of 77%, CI: 0.55 to 0.99; specificity of 100%) in differentiating between the DLD

[2] CI=95% confidence interval

[3] Although the confidence interval mathematically exceeds 1, the maximum value of sensitivity/specificity is 1.

group and the age-matched TD group, compared to the percent correct syllable method (sensitivity of 43%, CI: 0.17 to 0.69; specificity of 100%).

To summarize, the SR task has received support as a clinical marker for DLD across languages. How SR responses are scored has a direct bearing on the classification accuracy of the SR task (Leclercq et al., 2014; Pham & Ebert, 2020; Stokes et al., 2006). The error method is the most commonly used scoring system and consistently shows fair to good classification performance in different languages (Pham & Ebert, 2020; Redmond, 2005; Stokes et al., 2006; Wiig et al., 2013). The binary method is the simplest and yielded the highest sensitivity in Vietnamese (Pham & Ebert, 2020) and the best overall classification results in French (Leclercq et al., 2014). For a newly established SR task, these results highlighted the importance of empirically testing the most effective scoring method.

A Sentence Repetition Task in Mandarin

To the best of our knowledge, the utility of the SR task in Mandarin is yet to be established. The investigation of the SR task in Mandarin, a typologically distinct language from Indo-European languages, can further support this task's utility to differentiate individuals with and without DLD across languages and contribute to the understanding of the underlying deficits of DLD (Pham & Ebert, 2020). In addition, Chinese has 873 million native speakers and 178 million second language speakers all over the world, and Mandarin speakers constitute the majority of this population (Gordon, 2005). Following the logic that DLD is affected by genetic components and should be equally prevalent across languages and countries (Armon-Lotem et al., 2015; Rice, 2013), there are approximately 5 million 4-9 years old children in China that are estimated to have DLD (Sheng et al., 2020). The development of a Mandarin SR task as a screening tool is thus clinically significant to facilitate the early identification of DLD in Mandarin-speaking children.

There is a substantial literature on the manifestations of DLD in Mandarin in the preschool to early elementary school age period (Sheng et al., in preparation) that could guide the design of stimuli for a novel SR task. Specifically, this literature highlights several structures that are well-established in typically-developing children but present considerable challenges to children with DLD, including passives, classifiers, and aspect markers. Classifiers in Mandarin modify nouns that share the same properties in terms of shape or other dimensions (Lin & Bever, 2010), and they are mandatory when adding numerals to nouns (e.g., san1 *zhi1* gou3; three *classifier* dog). Passive sentences follow non-canonical word order and individuals with DLD across languages exhibit difficulty in this structure (Leonard, 2014). Zeng et al. (2018) found that children with DLD performed significantly lower than their TD age-matched peers on the comprehension and production of Mandarin passives. A narrative study also showed that Mandarin-speaking children with DLD produced significantly fewer passive sentences compared to their TD peers (Hao et al., 2018). In addition, Hao et al. (2018) observed significantly fewer classifiers and aspect markers used by children with DLD than TD children. Weaknesses in aspect markers were also shown in He and Sun (2013), which found that Mandarin-speaking children with DLD performed significantly worse than age-matched TD children on an aspect marker production task.

The Current Study

The current paper reports two studies that respectively examined the concurrent criterion validity of the self-designed Mandarin Sentence Repetition Task (MSRT) against criterion language measures of narrative sampling (study 1), and the discriminant validity (i.e., classification accuracy) of the MSRT against clinical diagnosis of DLD based on pediatrician judgment and standardized test scores (study 2). In study 1, to establish that the MSRT can reflect children's language ability, we examined if MSRT scores were correlated with measures derived from children's narrative samples. Through collecting and examining functional language use at the discourse level, researchers and clinicians could gain valuable insights into an individual's language abilities in everyday communication (Spencer et al., 2020). The analysis of language samples yields a deep and comprehensive evaluation of children's knowledge in different linguistic domains, including syntax, vocabulary, and use of specific linguistic structures (e.g., Andreu et al., 2011; Boudreau, 2008). Establishing that this newly designed MSRT aligns with measures derived from narrative samples in TD children is a crucial step before moving forward to examine the classification accuracy values in study 2.

In study 1, we derived four different measures from narrative samples to evaluate their relationship with children's SR performance, including mean length of utterance (MLU), vocabulary diversity (VOCD), number of predicates, and a composite structural measure. Both MLU and VOCD are commonly used in language sample analysis, as general measures of grammar (Boudreau, 2008; Justice et al., 2010) and vocabulary (Altman et al., 2016; Rezzonico et al., 2015) respectively. We did not use the type-token ratio (TTR) measure because it is subject to the influence of sample length: longer samples may give lower TTR values (Richards, 1987; Tweedie & Baayen, 1998). As opposed to a single value of TTR, VOCD is more informative because it represents how TTR varies over a range of token size for each speaker (Richards & Malvern, 2000).

Number of predicates is another measure of syntactic elements (Eisenberg, 2020), which include both verbs and predicate adjectives in Mandarin (Thomson & Tao, 2010). Devoscovi and

Cristina Caselli (2007) found significant correlations between children's performance on an Italian SR task and the number of predicates (verbs only in Italian) children used in spontaneous language samples. Moreover, to examine whether production of specific linguistic elements in an imitation context is related to production of the same elements in a spontaneous context, we included a composite structural measure that evaluates children's use of the linguistic structures that are featured in the SR stimuli, which will be named in the methods section.

As scoring methods may directly impact the classification accuracy of SR tasks, we explored five commonly used scoring methods in study 1. The scoring system(s) that did not lead to ceiling/floor effects and demonstrated significant correlations with all narrative measures were retained in study 2. We recruited Mandarin-speaking children with and without DLD in study 2 and examined the classification accuracy of the self-designed MSRT to differentiate these two groups. Showing correlations with gold standard measures is insufficient in demonstrating the clinical utility of a task, as reflecting general language abilities in TD children does not equal to accurately identifying DLD on a child-by-child basis. We calculated sensitivity, specificity, and likelihood ratios to evaluate the task's classification accuracy. We used Receiver Operating Characteristic (ROC) analysis to generate the optimal cutoff scores for the practical use of MSRT as a clinical screening tool. Overall, we aimed to answer three research questions. Study 1 addressed the first two questions and study 2 addressed the last question.

1) Does the MSRT reflect TD Mandarin-speaking children's language ability, by showing significant correlations with four narrative measures: MLU, VOCD, number of predicates, and a composite structural measure?

2) Out of the five scoring methods, which method(s) are not subject to ceiling/floor effects and can reflect children's language competence by demonstrating significant relationships with the four narrative measures?
3) Is the MSRT able to differentiate Mandarin-speaking children with and without DLD?

Study 1

Method

Participants

Fifty-nine Mandarin-speaking preschoolers (30 males, 29 females) participated in study 1. All participants were Asians of Chinese ethnicity. Parents of the participating children signed an informed consent approved by the University of Delaware's Institutional Review Board. Children's age ranged from 45 to 77 months, and the mean age was 61.7 months. The participants were recruited from the same preschool in Nanjing, China. All children were typically-developing with no reported language, sensory, speech production, motor, or cognitive disorders according to parent reports. Children's nonverbal intelligence was measured using the *Primary Test of Nonverbal Intelligence* (PTONI) (Ehrler & McGhee, 2008), and the average standardized score was 125.1 (SD=17.2, range=85-149). Participants' caregivers filled out a questionnaire to report children's family background, general health condition and developmental history, and Mandarin exposure.

Children's family SES was collected through surveying their maternal education, using a five-point likert scale. Twenty-five percent of the parents had a master's degree or higher, 56% had a bachelor's degree, 17% completed some college, and 2% completed middle school or lower. The mean maternal education score was 4 (bachelor's degree) (SD=0.77). Participants'

Mandarin exposure was collected using a five-point likert scale question asking about the percent of the waking hours that the child spent hearing and speaking Mandarin. Children's average Mandarin exposure score was 4.63 (SD=0.75), with a score of 4 indicating 60-79% and a 5 indicating 80-100%. Other than Mandarin, children were either exposed to dialects that are mutually intelligible with Mandarin (e.g., Henan and Nanjing dialects) and/or were exposed to other languages such as English (n = 40) and German (n = 1). It is worth noting that previous studies of Mandarin-speaking children's language development rarely if ever reported children's Mandarin exposure (Sheng et al., in preparation). Researchers likely assumed the monolingual status of their sample because of the prestige of Mandarin in the Chinese society: Mandarin is the only official language, the language of media, and the language of instruction at schools (Dong, 2010). Though the average amount of Mandarin exposure of the current sample seems low, given the similarity in recruitment approaches, we believe the sample is comparable to samples in previous studies of Mandarin language development, and is representative of the language exposure patterns of preschool age children in mainland China.

Test Materials

Sentence Repetition Task

In keeping with previous studies (Redmond, 2005; Stokes et al., 2006) and the review on the manifestations of DLD in Mandarin, we included eight sentences with passives and eight sentences with aspect markers as core elements. Each sentence type comprises eight test sentences with length ranging from 13 to 15 characters. Ninety percent of the nouns, verbs, and adjectives included in the sentences are early acquired lexical items, which appear in children's production as early as 17 months of age according to the Chinese Communication Development Inventory (CDI; Tardif & Fletcher, 2008). More challenging but age-appropriate words (ten percent, e.g., 美味, mei3-wei4, delicious; 批评, pi1-ping2, criticize; 偷走, tou1-zou3, steal) were included as well to increase the task's ability of revealing individual difference and avoid ceiling effect in TD children. The passive sentences constitute the elements of patient, "bei4" (passive marker), agent, and verb (see 1 for an example). The aspect marker sentences contain subject, verb, aspect marker, and object (see 2 for an example). Eleven sentences include classifiers (bolded in Appendix A), and four sentences contain embedded relative clauses (underlined in Appendix A). Relative clauses are featured in the stimuli to avoid ceiling effect in TD children, as they are structurally complex and acquired late developmentally (He et al., 2017; Sung et al., 2016). All sentences were pre-recorded by a female native speaker of Mandarin Chinese.

1）那只白色的小狗被妈妈抱走了。

Na4 zhi1 bai2 se4 de xiao3 gou3 bei4 ma1 ma1 bao4 zou3 le.

That CL white LP dog PASS mother carry away(RP) SFP

That white dog was carried away by the mother.

2） 小老鼠吃了一块美味的巧克力。

Xiao3 lao3 shu3 chi1 le yi1 kuai4 mei3 wei4 de qiao3 ke1 li4.

Little mouse ate PerM one CL delicious LP chocolate.

The little mouse ate a piece of delicious chocolate.

*CL = classifier; SFP = sentence final particle; PerM = perfective aspect marker; PASS = passive marker; LP = linking particle; RP = resultative particle

Narrative Task

Children completed the Mandarin version of the Multilingual Assessment Instrument for Narratives (MAIN, Gagarina et al., 2012; Luo et al., 2020). The MAIN was designed based on a series of pilot studies with more than 500 monolingual and bilingual children in 17 languages and 14 language pairs, and the stimulus pictures and scripts were carefully constructed to elicit narrative samples from children with diverse cultural, linguistic, and socio-economic backgrounds (Gagarina et al., 2012). The MAIN encompasses four stories, with two stories assigned to the story-telling format (*dog* and *cat*) and two stories assigned to the story-retelling format (*baby bird* and *baby goat*). All stories are parallel regarding cognitive and linguistic complexity, cultural appropriateness, and test robustness (Gagarina et al., 2012). Each story has a setting and episode structures that can capture the universal organizational pattern of stories. The MAIN has been used in 15 different languages as a language assessment tool (e.g., Dutch: Blom et al., 2020; Greek: Tsimpli et al., 2020; Italian: Levorato & Roch, 2020; Cantonese: Chan et al., 2020; Mandarin: Sheng et al., 2020).

Procedures

Sentence Repetition Task

Each child was assessed individually in a private room at their school. The task was administered using a PowerPoint presentation on a computer. The experimenter sat at a table next to the child and started with playing the instructions on the computer. The instruction was: "Let's play an imitation game on the computer. The computer will say some sentences. Your job is to listen carefully to what the computer says and repeat the sentences. Please remember, you have to say exactly the same thing as the computer. Let's practice!". Following the instructions, two practice items were administered one at a time to ensure that children understood the task. For the practice items, the experimenter guided the child to repeat the sentences and corrected them if they did not repeat verbatim. The test phase began once the child had successfully repeated the practice sentences word for word. A total of 16 test sentences were then presented one at a time at 65dB SPL using the computer's built-in speaker. The experimenter pressed a button on the computer to play the next sentence once the child had responded. The experimenter was instructed not to interfere with children's performance in any means besides providing general encouragement for children to continue during the actual test. The SR sessions were recorded using a Philips VTR5100 voice recorder with the noise reduction function. Transcriptions were completed later based on the recordings.

Narrative Task

Two tasks were administered in this study: first a story-tell task and then a story-retell task. The child was presented with three envelopes containing the same story inside (*baby goat* or *baby bird*) and was told that each envelope contained a different story. The experimenter pretended that they did not know which story the child would choose to create a more interactive environment and motivate the child to tell the story in detail. After the child had made their decision, the experimenter held the pictures facing the child to give them an overview of the story. When the child was ready to tell the story, the experimenter presented them with two pictures at a time. At the end of each two pictures, the experimenter encouraged the child to say more using prompts like "What else?", "Can you tell me more?", or "Is that all?". In the retell task, the child was presented with another three envelopes containing the same story (*dog* or *cat*). After the child made their decision, the examiner told a model story using the script provided by Luo et al. (2020). Subsequent to the demonstration, the child was asked to retell the story with the aid of the pictures. They were presented with two pictures at a time to ensure that they followed the story in sequence. The same prompts were given at the end of each two pictures.

The narrative sessions were recorded using the same voice recorder. Transcriptions were completed later.

Transcription and Scoring

Sentence Repetition Task

Responses to each sentence were transcribed verbatim using a Microsoft excel worksheet. We explored five potential scoring methods based on the literature. The first two followed the error method (Conti-Ramsden et al., 2001; Stokes et al., 2006; Wiig et al., 2013) and are respectively referred to as error method-word and error method-syllable. In Mandarin, a syllable represents one character, and a word may constitute one (e.g., 我, wo3, “me”) or more syllables/characters (e.g., 公园, gong1-yuan2, “park”). Two error methods are explored because words and syllables are both potential basic units of grammar in Mandarin (Duanmu, 2016) and the counting unit (word vs. syllable) may alter the number of errors. For example, a substitution of “蛋糕, dan4-gao1, cake” with “饼干, bing3-gan1, cookie” would be counted as one error in error method–word and two errors in error method–syllable. In both error methods, a score of three is given to completely accurate repetitions, a score of two is given if one error is found in a response, a score of one is given to a response that contained two to three errors, and a score of zero is assigned to a response that contained four errors or more. Any deviation in words or syllables from the target sentence (e.g., substitution, deletion, addition) was counted as one error. In both error methods, the score for each child was calculated by dividing the child’s score by the total possible score of 48 (3 x 16 sentences). Mazes (e.g., filled pauses, revisions) were not included as errors in the two error methods.

The third method is the binary scoring method (Newcomer & Hammill, 2019; Rispens, 2004), in which only a completely correct repetition is given a score of 1. Responses with any deviations from the target were scored as 0. The percentage of completely correct repetitions (out of 16) were calculated for each child. The fourth method is the core element method, also known as grammatical scoring (Pham & Ebert, 2020; Stokes et al., 2006). Using this method, a response receives a score of 1 if it contains all the core elements of the two stimulus types. For the passive sentences, the core elements are bei4 + agent + verb, whereas for the aspect marker sentences, the core elements are verb + aspect marker + noun. The last method is the percent of correct syllable method (Stokes et al., 2006), which calculates the percentage of correct syllables. Additions and transpositions were not penalized. As speech production errors may act as a confound in a language task and negatively affect children's scores, we did not take any points off for clear speech production errors (e.g., pronouncing /t^h/ as /t/; 糖, tang2 -> dang2). The illustration of the five scoring methods for an example response is presented in the supplementary material.

Narrative Task

Children's narrative samples were transcribed into Chinese characters using the Codes for Human Analysis of Transcripts (CHAT; MacWhinney, B., 2000). The transcriptions were analyzed using the Computerized Language Analysis (CLAN; MacWhinney, B., 2000). The following measures were calculated from the transcriptions of both the tell and retell samples:

1) Mean Length of Utterance (MLU) and vocabulary diversity (VOCD)

   All transcriptions were segmented into independent clauses which constitute the basis of utterance measures (Sheng et al., 2020). The utterances were first segmented into words using the “*Chinese online word segmentation system*” (http://ckipsvr.iis.sinica.edu.tw) and then checked manually. MLU-word and VOCD were generated using the CLAN system.

2) Number of predicates: the total number of verbs and predicate adjectives in children’s telling and retelling samples were manually coded and calculated. Verbs include both action verbs (e.g., 吃, chi1, eat; 看, kan4, look; 跑, pao3, run) and modal auxiliary verbs (e.g., 可以, ke2-yi3, can; 要, yao4, want/will). Predicate adjectives are the adjectives used in sentences without verbs.

3) Structural composite: Children’s correct use of classifiers, aspect markers, passives and relative clauses was manually coded and calculated. The general classifier *ge4* was not included in the count of classifiers. Every unique combination of classifier + noun was counted. For example, 一只羊 (one *CL-zhi1* sheep) and 一只鸟 (one *CL-zhi1* bird) were counted as two different classifier uses. Two types of aspect markers were coded and counted in this measure: the progressive markers (zai4 and zhe) and the perfective markers (le and guo4). There are two different uses of le in Chinese: one is a genuine perfective marker indicating the perfective aspect, and the other is a sentence-final particle that marks the reported event or situation as “relevant” to the context (Li & Thompson, 1989; Wang & Sun, 2015). In this analysis, only the correct use of le as a perfective aspect marker was counted. Total number of grammatical passive and relative clause sentences was calculated as well. The structural composite score was derived by adding up the number of correct uses of the four structures in children’s narrative samples (tell and retell).

Reliability

The first author of this paper transcribed and coded all participants' SR responses and narrative samples. Twenty percent of the transcriptions, scoring and coding were completed independently by a native Mandarin-speaking research assistant. Reliability of the transcriptions for the SR and narrative tasks was evaluated on the basis of Chinese characters. Each Chinese character corresponds to a syllable/morpheme that can be an independent word or part of a word (Wiedenhof, 2015). Reliability was calculated using the number of consistent characters divided by the number of total characters for the MSRT. For the narrative task, the reliability was calculated by averaging the following two values: 1) number of consistent characters divided by total number of characters transcribed by transcriber 1; 2) number of consistent characters divided by total number of characters transcribed by transcriber 2. The transcription reliability was 96% for the MSRT and 94% for the narrative task. Reliability of SR coding was calculated using the number of consistent scorings divided by the total number of scorings, yielding a reliability of 92%. For the coding of narrative samples, the mean agreements on utterance and word segmentation were 93% and 91% respectively. Utterance and word segmentation formed the basis for the automatic calculation of MLU and VOCD using the CLAN software. The number of predicates, aspect markers, classifiers, passives, and relative clauses in children's narrative samples was coded twice for 20% of the participants, yielding inter-rater agreements of 98%, 95%, 98%, 96%, and 96% respectively. All inconsistencies were discussed between the two coders until consensus was reached.

Results

We first conducted the Shapiro-Wilk test to check the normality of all variables and determined the appropriate analysis approach to examine the relationship between the MSRT and the narrative measures. The SR scores using the error method-word and binary scoring, MLU, VOCD, the structural composite, and number of predicates were normally distributed. SR scores using other scoring methods were not normally distributed. Despite the fact that some variables were normally distributed, we employed the non-parametric Spearman's rank order correlation for all pairs to allow reasonable comparison among the correlational outcomes for the five scoring methods. Though outliers were present, the Spearman's rank order correlation does not require outliers to be removed.

Participants' SR scores are presented in Figure 1. Both the core element scoring and the percent of correct syllable scoring led to a ceiling effect. Using the core element scoring, over half of the children (N=33) scored 100% on the MSRT; and using the percent of correct syllable scoring, over 70% of the children (N=45) scored between 90% and 100%. Table 1 shows the results of the Spearman rank correlations. The scatterplots of the relationship between children's SR scores and performance on the narrative measures were presented in the supplementary material. The *p* level was corrected for multiple correlations using the Bonferroni's correction (corrected significant level: $p < .003$). Four of the five scoring methods demonstrated significant correlations with all narrative measures. The core element scoring did not correlate with MLU, and the number of predicates in narratives.

Insert Figure 1 about here

Insert Table 1 about here

Interim Discussion

In study 1, we validated the self-designed Mandarin SR task (MSRT) against four language measures derived from children's narrative samples (MLU, VOCD, number of predicates, structural composite). The results demonstrated that using four out of the five scoring methods, children's SR performance significantly correlated with all narrative measures, indicating that the MSRT can reflect Mandarin-speaking children's language abilities. The logical next step is to further evaluate the classification accuracy of this task in differentiating between Mandarin-speaking children with and without DLD.

Study 1 results could further guide the selection of scoring methods in the classification accuracy study. Specifically, the core element scoring could be excluded as it correlated with only two of the four narrative measures and showed a ceiling effect. Although the percent of correct syllable scoring did show significant correlations with the narrative measures, it also led to a ceiling effect, which is not desirable in test development. The two error methods and the binary method appeared to be superior to the other methods because they showed significant correlations with narrative measures, and they elicited a wide range of performance. The two error methods showed comparable results but differed in ease of use. The judgment of syllables in Mandarin is more straightforward than that of words, as each syllable is equivalent to one character and each word may contain one or more syllables. The definition of word in Mandarin can be controversial (Li & Thompson, 1989). For example, Jespersen (1922) concluded that Mandarin words are essentially monosyllabic, while Kennedy (1951) and Lin (1952) argued that most Chinese words occur in disyllabic forms. Other researchers even suggest that wordhood is nonexistent in Chinese languages (Chao & Yang, 1947; Chéng, 2003). Accurate word segmentation requires a combination of automated segmentation software and manual correction, which adds considerable time to the scoring process and requires substantial linguistic knowledge from researchers and clinicians. We therefore further explored the error method in syllable and the binary method in study 2.

Study 2

Method

Participants

Sixty-nine Mandarin-speaking children between the ages of 4;0 and 5;11 were recruited and tested. All but one participants were Asians of Chinese ethnicity. One child was of mixed Chinese and Japanese ethnicity. Parents of the participants signed an informed consent approved by the Research Ethics Board of the Shanghai Children's Medical Center. Participant recruitment was conducted in three phases. In phase one, a one-gate design was applied, with both groups recruited simultaneously from the same site, the Developmental and Behavioral Pediatrics Department at the Shanghai Children's Medical Center. Thirty-eight children who visited the clinic and were screened as having no physical or neurological impairments participated from March 2019 to December 2019. Further classification of this group yielded five TD children in phase one. Given the inefficiency of recruiting TD children from an outpatient clinic setting, in phase 2, we recruited TD children from local communities using word of mouth and advertising on a popular social media platform in China, WeChat. Ten TD children were recruited and tested in two weeks. Phase 2 was paused in January 2020 because of the outbreak of the Coronavirus Disease in China. Recruitment for the TD group resumed in June 2020, and 21 children were recruited and tested within a month.

Participating children first received a hearing screening from a pediatrician using a portable audiometer. Children who have normal hearing then received a clinical screening conducted by a developmental and behavioral pediatrician with over 30 years of experience to rule out other physical and neurological impairments. The decisions were made based on the standards described in the *Diagnostic and Statistical Manual of Mental Disorders-V* (DSM-V; APA, 2013) and children's medical history. Children who had normal hearing, normal or corrected-to-normal vision, and no diagnosis of neurological disorder, speech production deficits, autism, genetic disorder, or cerebral palsy were invited to complete the *Wechsler Preschool and Primary Scale of Intelligence* – Revised (WPPSI-R; Wechsler, 1989). The WPPSI-R is a standardized intelligence scale designed for children aged 3;0 to 7;3. A Chinese version is available (Chen & Chen, 2000), which was normed on 900 Taiwanese children. To be included in study 2, children needed to score over 80 on the WPPSI-R performance scale (.89 of Cronbach's alpha and test-retest reliability; Chen & Chen, 2000), to demonstrate normal nonverbal intelligence. Four children were excluded from the study because they did not meet the inclusion criterion for performance IQ.

Qualified children were then invited to complete the *Diagnostic Receptive and Expressive Assessment of Mandarin* (DREAM; Ning et al., 2014), which was used for language status classification. The DREAM is a standardized, norm-referenced oral language assessment for Mandarin-speaking children ages 2;6 to 7;11. The test provides one total score and four component scores: expressive language, receptive language, semantics, and syntax. DREAM achieved high test-retest reliability (r=.85) and good external validity by demonstrating significant correlations with spontaneous language indices (e.g., sentence complexity and vocabulary diversity) and narrative indices (e.g., use of mental verbs and connectives) (Liu et al., 2017). When validated against a combination of pediatricians' judgment and spontaneous language samples, a cutoff score of 80 on any one of the DREAM components yielded a sensitivity of 95% and a specificity of 82% in differentiating children with and without DLD (Liu et al., 2017). We followed these empirically-derived guidelines regarding the use of the DREAM test scores and included children who had at least one component standard score at or below 80 (the 10^{th} percentile or 1.3 SD below the mean) in the DLD group.

Of the 69 children recruited in study 2, 27 children met the criteria as having DLD. One child with comorbid ADHD was included in the DLD group, as the most recent DLD definition does not exclude children with ADHD (Bishop et al., 2017). Eight children were excluded as their parents did not complete the parent questionnaire to report family background and language environment. One child was excluded as one of his/her parents is a native Japanese speaker. We then attempted to select an age-matched TD peer for each of the remaining 18 children with DLD. An eligible TD match needed to 1) have all DREAM component scores higher than 80; 2) be within six months of age of the child with DLD; 3) have a Mandarin exposure score within ± 1 from the child with DLD; and 4) have a maternal education score within ± 1 from the child with DLD. We were able to find TD matches for 16 children with DLD. As shown in Table 2, the TD and DLD groups did not differ significantly on age, maternal education, and Mandarin exposure. The DLD group showed significantly lower nonverbal IQ scores and DREAM total scores than the TD group.

Insert Table 2 about here

Procedures

Each child participated in the study in a quiet assessment room in the Developmental and Behavioral Pediatrics Department of the Shanghai Children's Medical Center. A trained native Mandarin-speaking research assistant administered the MSRT to the child as part of a larger battery of language and cognitive tests. The MSRT was given in the middle of a larger language assessment battery and the administration followed the same procedures as described in study 1. The error method in syllable and the binary method were used to score children's responses.

Reliability

The first author of this paper transcribed and scored all participants' responses. A second native Mandarin-speaking trained research assistant transcribed and scored 20% of the data independently to examine reliability. During transcription and scoring, both the first author and the research assistant were blinded to the grouping status of the children to avoid potential biases. Transcription reliability was calculated by dividing the number of consistent characters by the number of total characters in the sentences, yielding an inter-rater reliability of 96%. Scoring reliability was calculated by dividing the number of consistent scorings by the total number of scorings, yielding an overall inter-rater reliability of 95%. Disagreements were resolved by reaching consensus between the two coders.

Analyses

We first compared the DLD and TD groups' performance on the MSRT to examine whether the two groups differed on this task. As the SR scores using the error method in syllable for both groups and the SR scores using the binary method for the DLD group were not normally distributed, Mann-Whitney U tests were conducted to compare the two groups' MSRT scores. Receiver Operating Characteristic (ROC) curves were generated using SPSS v.26 to determine the optimal cutoff point, sensitivity, and specificity for the MSRT. The ROC curve is a graph which plots the sensitivity and specificity of a binary classification system as the discrimination threshold (cutoff) varies (Fluss et al., 2005). Each ROC curve generates a value for the area under the curve (AUC), which represents an overall estimate of the task's accuracy in classifying individuals as with and without DLD. The AUC values are interpreted following the guidelines in Swets et al. (2000): values between 0.90-1.0 are considered "excellent"; values between 0.80-0.90 are considered "good", values between 0.70-0.80 are considered "fair", and values lower than 0.70 are considered "poor". Following Redmond et al. (2019), the optimal cutoff points on the ROC curves were identified using the Youden index (*J*) (Youden, 1950). The Youden index *J* value captures the performance of a diagnostic test, and it is calculated following the formula of *J* = sensitivity + specificity -1. A *J* value of 0 indicates complete overlap between the affected and unaffected groups and suggests that this classification task is useless. A *J* value of 1 indicates that the task could completely separate affected and unaffected groups. Therefore, the optimal cutoff point should be associated with the maximum *J* value.

Once the optimal cutoff points and their associated sensitivity and specificity were identified, the positive (LR+) and negative (LR-) likelihood ratios were calculated from sensitivity and specificity values. LR+ and LR- respectively represents the probability that a person with the condition testing positive for the condition and the probability that a person without the condition testing negative for the condition. The likelihood ratios were calculated using the following formula: LR+ = sensitivity / (1-specificity); LR- = (1-sensitivity)/specificity. We followed the guidelines specified in the introduction to interpret the sensitivity and specificity values (Plante & Vance, 1994). Likelihood ratios were interpreted following Dollaghan (2007): LR+ values between 3 and 10 indicate moderate positivity or a suggestive level of clinical informativeness for identifying DLD; LR+ values at or above 10 are confirmatory and clinically informative; LR- values between .1 and .2 indicate moderate negativity; and LR- values at or below .1 are exclusionary and indicates high confidence in ruling out a child with DLD.

As we intended for the MSRT to be used as a screening tool, greater emphasis is placed on achieving high sensitivity (US Preventive Services Task Force, 2006). Children who perform poorly on a screening tool would receive more comprehensive language assessments to confirm their status. Therefore, this process would tolerate some false positives (inaccurately identify a

TD child as having DLD), as the misidentifications will be cleared up through further evaluation. On the other hand, false negatives (inaccurately identify a child with DLD as TD) are not desirable as they filter out children with DLD and stop them from receiving further evaluation and intervention. A high sensitivity value indicates that there are few false negative results, thus fewer cases of DLD are missed. Therefore, a good screening task would desire high sensitivity values and may tolerate lower specificity values.

Results

Group comparisons

Figure 2 presents the two groups' SR scores. The middle line in each vertical box represents the median, and the upper and lower lines represent the third and first quartile of the data respectively. Using the error method in syllable, the DLD group received a mean score of 0.19 (SD=0.19) and the TD group received a mean score of 0.76 (SD=0.16). Using the binary method, the DLD group received a mean score of 0.06 (SD=0.09) and the TD group received a mean score of 0.62 (SD=0.21). Mann-Whitney U tests revealed that the TD group achieved significantly higher scores compared to the DLD group using both scoring methods (error method in syllable: W=140, p<.001, Cohen's d=3.25; binary method: W=141, p<.001, Cohen's d=3.47).

Classification accuracy

The MSRT achieved AUC values of .984 and .982 using error method in syllable and binary method, both demonstrating excellent classification accuracy. Using the error method in syllable, an optimal cutoff score of .63 yielded sensitivity and specificity values of 100% and 87.5%. Using the binary method, an optimal cutoff score of .41 yielded the same sensitivity and specificity values of 100% and 87.5%. The positive likelihood ratio is 8.0 using both scoring methods, which demonstrate moderate classification of children with DLD. The negative likelihood ratio is 0 using both scoring methods, which suggests high confidence in ruling out a child with DLD.

Discussion

The current study aimed to design and validate a Mandarin sentence repetition task. In study 1, we investigated the criterion validity of the MSRT by examining correlations between TD children's performance on the MSRT and benchmark measures of language skills based on a narrative task. Four narrative measures were derived, including two syntactic measures (MLU and number of predicates), a vocabulary measure (VOCD), and a composite measure of four linguistic structures (aspect markers, classifiers, passives, relative clauses). In addition, as different scoring methods yielded distinct performance of SR tasks (Pham & Ebert, 2020; Redmond, 2005; Stokes et al., 2006), we explored five potential scoring systems to determine the best scoring method for the MSRT. In study 2, we calculated the sensitivity, specificity, and LRs associated with the optimal cutoff score to examine whether the MSRT could accurately differentiate between Mandarin-speaking preschoolers with and without DLD.

Study 1 showed that the MSRT is a valid measure to evaluate Mandarin-speaking preschoolers' language ability. The two error methods, the binary method, and the percent of correct syllable method significantly correlated with all validation measures, illustrating that the MSRT in general could reflect children's language abilities as measured through narrative sampling. However, scores using the percent of correct syllable method was right skewed and resulted in a ceiling effect. This ceiling effect could be attributed to that the percent correct syllable method did not consider the addition or the transposition of the syllables/words. In some cases, a syllable corresponds to a stand-alone word in Mandarin, and word order is an important consideration as children with DLD demonstrate difficulties with following the correct word orders (Hansson & Nettelbladt, 1995). The core element scoring, which only considered the accuracy of predefined grammatical targets, did not show significant correlations with MLU and number of predicates. The core element scoring does not consider other elements in the sentences except for the target structures, thus providing only a partial picture of children's language abilities. In addition, the core element scoring resulted in a ceiling effect with over 50% of children scoring 100% accurate on the task, lowering its ability to reveal individual differences among children.

Study 2 showed that the MSRT can adequately differentiate Mandarin-speaking children with and without DLD using both the error method in syllable and the binary scoring method. Additional investigation in a new sample is needed to further compare the classification accuracy associated with the two scoring methods. The AUCs derived from the ROC analysis demonstrated excellent classification accuracy and the classification accuracy values all indicate acceptable to good classification power. The 100% sensitivity using both scoring methods further supported the MSRT as a language screening task for DLD. This study verifies the utility of SR tasks to differentiate between children with and without DLD in a language that is typologically distinct from the most studied Indo-European languages (e.g., Conti-Ramsden et al., 2001; Leclercq et al., 2014; Redmond, 2005). In addition, our finding showed that the error method in syllable and the binary method yielded the same classification accuracy in a Mandarin SR task, which differs from the findings in other Asian languages wherein the error method achieved higher overall classification accuracy compared to other scoring methods, including the binary method (Cantonese: Stokes et al., 2006; Vietnamese: Pham & Ebert, 2020).

The high sensitivity of the MSRT could be attributed to our stimulus design that took into consideration known areas of learning difficulties in Mandarin, including classifiers – noun-modifying morphemes that are semantically complex (Hao et al., 2021), and two grammatical features – passives (Zeng et al., 2018) and aspect markers (He & Sun, 2013). Errors on these vulnerable structures were frequently observed in the responses produced by children with DLD. For example, substitutions of specific classifiers with the general classifier *ge* (e.g., one *CL-kuai* chocolate -> one *CL-ge* chocolate) and omissions of aspect markers (ZAI and LE) were quite common. In addition, children with DLD changed passive sentences (e.g., *The wolf is defeated by the smart goat.*) to either simple sentences (e.g., *The wolf defeated the smart goat.*) or BA-sentences in Mandarin (e.g., *The wolf BA goat defeated.*). Future studies that more closely examine the error patterns in children with DLD could further shed light on the linguistic manifestations of DLD in Mandarin.

Clinical Utility of the MSRT

The design and validation of the MSRT remediates the paucity of Mandarin language evaluation tools by providing clinicians and researchers with a quick screening tool to differentiate children with and without DLD at an initial stage. As described earlier in the paper, a high sensitivity value is desirable for a good screening task. The current MSRT is thus a promising screening tool as the sensitivity is 100% and the specificity is 87.5% using both the error (syllable) and binary methods. In addition, a screening task needs to be time-efficient so that it can be given to a large number of individuals. For a vastly under-diagnosed disorder such as DLD, this is especially important because the uncovering of the many hidden cases may require universal screening. From our data collection experience, administering the MSRT takes about six minutes and transcription and scoring using the error method take about eight minutes, adding up to approximately 14 minutes to screen one child. When using the binary method, the scoring can be completed online without transcriptions, which makes it around six minutes to screen one child. The binary method may be selected when time is of the essence. The error method in syllable may be selected when the examiner desires more in-depth information on the child's error patterns. Moreover, the task does not require intensive training or considerable linguistic expertise and can be administered and scored by classroom teachers or nurses. This short and valid screening test will facilitate the allocation of limited resources to identify children who are in need of further assessments and timely intervention.

It is important to use the optimal scoring methods and cutoff points in the real-life application of MSRT as a screening tool for DLD. The sensitivity and specificity values presented in this paper were associated with the reported optimal cutoff point (accuracy of 0.63 for error method, accuracy of 0.41 for binary method). Using a different cutoff may negatively impact the classification accuracy of this task. In addition, we recruited the DLD group in a hospital setting, and the need of medical assessment/consultation suggested parent concerns of children's language and behavior. We recommend using a combination of the MSRT and a measure of parent concern when viable, which may yield more accurate classification results when using the MSRT as a universal screening tool.

Limitations and future directions

We employed a two-gate design instead of the more desired one-gate design in study 2. In a one-gate design, all participants are recruited from a single population, whereas in a two-gate design, affected and unaffected groups are recruited from separate populations. Although only two of the 13 English diagnostic studies reviewed in Pawłowska (2014) used the one-gate design in participant recruitment, the author emphasized the importance of a one-gate design in diagnostic accuracy studies to avoid the influence from the fundamental differences across populations. In study 2, we started with a one-gate design by recruiting both groups from a hospital's outpatient clinic. However, this proved inefficient as we recruited only five TD children during a period of nine months. We had to change course and recruited additional TD children from the local community. Potential differences in sample characteristics may partly contribute to the good classification accuracy of MSRT presented in this study.

We used an exploratory sample to examine the classification utility of the MSRT and the corresponding cutoff score. Future studies should replicate the results with a confirmatory sample of children with and without DLD to verify the classification power of this task and further test the utility of the two scoring methods. A complete list of the sentences is provided in appendix A. To achieve the one-gate design in participant recruitment, future studies could either extend over a longer period in a hospital setting or carry out larger-scale screening in schools to ensure that the participants are from the same population. Moreover, future studies could conduct qualitative analysis of repetition response and compare children with and without DLD. Woon et al. (2014) suggested that SR tasks are good candidates for qualitative evaluations and the errors identified in SR tasks can help us understand the underlying deficits and particular weaknesses associated with DLD in Mandarin-speaking populations.

Acknowledgment

The authors wish to thank the participating families for volunteering their time, the research assistants at Nanjing Normal University and the clinician in the Developmental and Behavioral Pediatrics Department in the Shanghai Children's Medical Center for administering the tasks, Huanhuan Shi and Lue Shen for helping with the reliability check of the transcriptions and coding, and Dr. Pumpki Lei Su for providing insightful comments to an earlier draft of the paper.

References

Alloway, T. P., & Gathercole, S. (2005). Working memory and short-term sentence recall in young children. *European Journal of Cognitive Psychology*, *17*(2), 207-220.

Altman, C., Armon-Lotem, S., Fichman, S., & Walters, J. (2016). Macrostructure, microstructure, and mental state terms in the narratives of English-Hebrew bilingual preschool children with and without specific language impairment. *Applied Psycholinguistics*, *37*(1), 165.

American Psychiatric Association, & American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders: DSM-5. *Arlington, VA*.

Andreu, L., Sanz-Torrent, M., Guàrdia Olmos, J., & Macwhinney, B. (2011). Narrative comprehension and production in children with SLI: An eye movement study. *Clinical linguistics & phonetics*, *25*(9), 767-783.

Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research*, *52*(4), 899-914.

Armon-Lotem, S., de Jong, J., & Meir, N. (Eds.). (2015). *Assessing multilingual children: Disentangling bilingualism from language impairment*. Multilingual matters.

Armon-Lotem, S., & Meir, N. (2016). Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: evidence from Russian and Hebrew. *International journal of language & communication disorders*, *51*(6), 715-731.

Baddeley, A. (2000). Working memory and language processing. *Benjamins translation library*, *40*, 1-16.

Bedore, L. M., & Leonard, L. B. (2001). Grammatical morphology deficits in Spanish-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research*.

Bishop, D. V. (1997). Cognitive neuropsychology and developmental disorders: Uncomfortable bedfellows. *The Quarterly Journal of Experimental Psychology Section A*, *50*(4), 899-923.

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Catalise Consortium. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLOS one*, *11*(7), e0158753.

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., Catalise-2 Consortium, Adams, C., … & Boyle, C. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, *58*(10), 1068-1080.

Blom, E., Boerma, T., & de Jong, J. (2020). Multilingual Assessment Instrument for Narratives (MAIN) adapted for use in Dutch. *ZAS Papers in Linguistics*, *64*, 51-56.

Bortolini, U., Caselli, M. C., & Leonard, L. B. (1997). Grammatical deficits in Italian-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *40*(4), 809-820.

Boudreau, D. (2008). Narrative abilities: Advances in research and implications for clinical practice. *Topics in Language Disorders*, *28*(2), 99-114.

Chan, A., Cheng, K., Kan, R., Wong, A. M. Y., Fung, R., Wong, J., ... & Gagarina, N. (2020). The Multilingual Assessment Instrument for Narratives (MAIN): Adding Cantonese to MAIN. *ZAS Papers in Linguistics*, *64*, 23-29.

Chao, Y. R., & Yang, L. S. (1949). *Concise Dictionary of Spoken Chinese*, Cambridge, Mass.: Harvard University Press.

Chen, J. H., & Chen, H. Y. (2000). Manual for the Wechsler preschool and primary scale of intelligence-revised. *Taipei, Taiwan: Chinese Behavioral Science Corporation*.

Chéng Y. (2003).*汉语字基语法: 语素层造句的理论和实践* [A zì–based grammar of Chinese: theory and practice in building sentences with morphemes], Shanghai: Fudan University Press.

Christensen, R. V. (2019). Sentence Repetition: A Clinical Marker for Developmental Language Disorder in Danish. *Journal of Speech, Language and Hearing Research (Online)*, *62*(12), 4450-4463.

Clarke, M. G., & Leonard, L. B. (1996). Lexical comprehension and grammatical deficits in children with specific language impairment. *Journal of communication disorders*, *29*(2), 95-105.

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of child psychology and psychiatry*, *42*(6), 741-748.

Devescovi, A., & Cristina Caselli, M. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language & Communication Disorders*, *42*(2), 187-208.

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia research: have we reached the tipping point?. *Aphasiology, 32*(4), 459-464.

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Paul H Brookes Publishing.

Dong, J. (2010). The enregisterment of Putonghua in practice. *Language & Communication, 30*(4), 265-275.

Duanmu, S., & Dong, Y. (2016). Elastic words in Chinese. In *The Routledge Encyclopedia of the Chinese Language* (pp. 490-506). Routledge.

Ehrler, D. J., & McGhee, R. L. (2008). *PTONI: Primary test of nonverbal intelligence*. Austin, TX: Pro-Ed.

Eisenberg, S. L. (2020). Using general language performance measures to assess grammar learning. *Topics in Language Disorders*, *40*(2), 135-148.

Finestack, L. H., & Satterlund, K. E. (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology*, *27*(4), 1329-1351.

Fleckstein, A., Prévost, P., Tuller, L., Sizaret, E., & Zebib, R. (2018). How to identify SLI in bilingual children: a study on sentence repetition in French. *Language Acquisition*, *25*(1), 85-101.

Fletcher, P., Leonard, L. B., Stokes, S. F., & Wong, A. M. Y. (2005). The expression of aspect in Cantonese-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research*.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *47*(4), 458-472.

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., … & Walters, J. (2012). MAIN: Multilingual assessment instrument for narratives. *ZAS papers in linguistics*, *56*, 155-155.

Gordon Jr, R. G. (2005). Ethnologue, languages of the world. http://www. *Ethnologue. Com/*

Hansson, K., & Nettelbladt, U. (1995). Grammatical characteristics of Swedish children with SLI. *Journal of Speech, Language, and Hearing Research*, *38*(3), 589-598.

Hao, Y., Sheng, L., Zhang, Y., Jiang, F., de Villiers, J., Lee, W., & Liu, X. L. (2018). A narrative evaluation of Mandarin-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research*, *61*(2), 345-359.

He, X., Sun, L. (2013). 汉语特殊性语言障碍儿童体标记“了”和“在”的产出研究 [The production of aspect markers “le” and “zai” by Mandarin-speaking children with specific language impairment]. *Journal of Foreign Language Education, 34*(2), 27-32.

He, W., Xu, N., & Ji, R. (2017). Effects of age and location in Chinese relative clauses processing. *Journal of psycholinguistic research*, *46*(5), 1067-1086.

Hwang, M. (2012). Sentence repetition as a clinical marker of specific language impairment in Korean-speaking preschool children. *Communication Sciences & Disorders*, *17*(1), 1-14.

Jessup, B., Ward, E., Cahill, L., & Keating, D. (2008). Teacher identification of speech and language impairment in kindergarten students using the Kindergarten Development Check. *International journal of speech-language pathology*, *10*(6), 449-459.

Jespersen, O. (1922). *Language: Its nature, development and origin* (Vol. 68). H. Holt.

Joanisse, M. F., & Seidenberg, M. S. (1998). Specific language impairment: A deficit in grammar or processing?. *Trends in cognitive sciences*, *2*(7), 240-247.

Johnston, J. R., & Kamhi, A. G. (1984). Syntactic and semantic aspects of the utterances of language-impaired children: The same can be less. *Merrill-Palmer Quarterly (1982-)*, 65-85.

Justice, L. M., Bowles, R., Pence, K., & Gosse, C. (2010). A scalable tool for assessing children’s language abilities within a narrative context: The NAP (Narrative Assessment Protocol). *Early Childhood Research Quarterly*, *25*(2), 218-234.

Kan, P. F., & Windsor, J. (2010). Word learning in children with primary language impairment: A meta-analysis.

Kennedy, G. A. (1951). The monosyllabic myth. *Journal of the American Oriental Society*, *71*(3), 161-166.

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental science*, *18*(1), 146-154.

Knox, E., & Conti-Ramsden, G. (2003). Bullying risks of 11-year-old children with specific language impairment (SLI): Does school placement matter?. *International Journal of Language & Communication Disorders*, *38*(1), 1-12.

Komeili, M., & Marshall, C. R. (2013). Sentence repetition as a measure of morphosyntax in monolingual and bilingual children. *Clinical linguistics & phonetics*, *27*(2), 152-162.

Laws, G., Briscoe, J., Ang, S. Y., Brown, H., Hermena, E., & Kapikian, A. (2015). Receptive vocabulary and semantic knowledge in children with SLI and children with Down syndrome. *Child Neuropsychology*, *21*(4), 490-508.

Leclercq, A. L., Quémart, P., Magis, D., & Maillart, C. (2014). The sentence repetition task: A powerful diagnostic tool for French children with specific language impairment. *Research in developmental disabilities*, *35*(12), 3423-3430.

Leonard, L. B., Miller, C., & Gerber, E. (1999). Grammatical morphology and the lexicon in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(3), 678-689.

Leonard, L. B., Kas, B., & Pléh, C. (2009). The use of tense and agreement by Hungarian-speaking children with language impairment.

Leonard, L. B. (2014). Specific language impairment across languages. *Child development perspectives*, *8*(1), 1-5.

Levorato, C., & Roch, M. (2020). Italian adaptation of the Multilingual Assessment Instrument for Narratives. *ZAS Papers in Linguistics*, *64*, 139-146.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar* (Vol. 3). Univ of California Press.

Lín, H. (1952).汉语是不是单音节语？[Is Chinese a monosyllabnic language?], *中国语文 11*, 6-11.

Liu, X. L., de Villiers, J., Ning, C., Rolfhus, E., Hutchings, T., Lee, W., … & Zhang, Y. W. (2017). Research to establish the validity, reliability, and clinical utility of a comprehensive language assessment of Mandarin. *Journal of Speech, Language, and Hearing Research*, *60*(3), 592-606.

Luo, J., Yang, W., Chan, A., Cheng, K., Kan, R., & Gagarina, N. (2020). The Multilingual Assessment Instrument for Narratives (MAIN): Adding Mandarin to MAIN. *ZAS Papers in Linguistics*, *64*, 159-162.

McGregor, K. K., Newman, R. M., Reilly, R. M., & Capone, N. C. (2002). Semantic representation and naming in children with specific language impairment.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Psychology Press.

Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. *Assessing multilingual children: Disentangling bilingualism from language impairment*, 95-124.

McArthur, G. M., Hogben, J. H., Edwards, V. T., Heath, S. M., & Mengler, E. D. (2000). On the "specifics" of specific reading disability and specific language impairment. *Journal of Child Psychology and Psychiatry*, *41*(7), 869-874.

Meir, N., Walters, J., & Armon-Lotem, S. (2016). Disentangling SLI and bilingualism using sentence repetition tasks: The impact of L1 and L2 properties. *International Journal of Bilingualism*, *20*(4), 421-452.

Newcomer, P., & Hammill, D. (2019). *Told-p: 5: Test of Language Development. Primary.* Austin, TX: Pro-Ed.

Ning, C. Y., Liu, X. L., & de Villiers, J. G. (2014). The diagnostic receptive and expressive assessment of Mandarin. *Dallas, TX: Bethel Hearing and Speaking Training Center*.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., … & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *Journal of Child Psychology and Psychiatry*, *57*(11), 1247-1257.

Paul, R., Norbury, C.F., & Gosse, C. (2017). *Language disorders from infancy through adolescence: Listening, speaking, reading, writing and communication* (5th ed.). St. Louis, MO: Mosby Elsevier.

Pawłowska, M. (2014). Evaluation of three proposed markers for language impairment in English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2261-2273.

Pham, G., & Ebert, K. D. (2020). Diagnostic Accuracy of Sentence Repetition and Nonword Repetition for Developmental Language Disorder in Vietnamese. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1521-1536.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15-24.

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure?. *International Journal of Language & Communication Disorders, 50*(1), 106-118

Redmond, S. M. (2005). Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology. *Clinical Linguistics & Phonetics*, *19*(2), 109-127.

Redmond, S. M., Ash, A. C., Christopulos, T. T., & Pfaff, T. (2019). Diagnostic accuracy of sentence recall and past tense measures for identifying children's language impairments. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2438-2454.

Rezzonico, S., Chen, X., Cleave, P. L., Greenberg, J., Hipfner-Boucher, K., Johnson, C. J., … & Girolametto, L. (2015). Oral narratives in monolingual and bilingual preschoolers with SLI. *International Journal of Language & Communication Disorders*, *50*(6), 830-841.

Rice, M. L. (2013). Language growth and genetics of specific language impairment. *International Journal of Speech-Language Pathology*, *15*(3), 223-233.

Richards, B. (1987). Type/token ratios: What do they really tell us?. *Journal of child language*, *14*(2), 201-209.

Richards, B., & Malvern, D. (2000). Measuring vocabulary richness in teenage learners of French.

Rispens, J. E. (2004). *Syntactic and phonological processing in developmental dyslexia*. Groningen: Rijksuniversiteit Groningen, Faculteit der Letteren.

Semel, E., Wiig, E., & Secord, W. (1994). *Clinical Evaluation of Language Fundamentals-Revised*. San Antonio, TX: The Psychological Corporation.

Sheng L. (2014). Semantic Development in Children with Language Impairments. *Encyclopedia of language development,* 534-538.

Sheng, L., & McGregor, K. K. (2010). Lexical–semantic organization in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 53*(1), 146-159.

Sheng, L., Shi, H., Wang, D., Hao, Y., & Zheng, L. (2020). Narrative Production in Mandarin-Speaking Children: Effects of Language Ability and Elicitation Method. *Journal of Speech, Language, and Hearing Research*, *63*(3), 774-792.

Sheng, L., Su, P.L., Wang, D., Yu, J., Lu, T.-H., Shen, L., Hao, Y., & Lam, B.P.W. (in preparation). Manifestations of developmental language disorder in Chinese children: A systematic review and meta-analysis.

Spencer, E., Bryant, L., & Colyvas, K. (2020). Minimizing variability in language sampling analysis: A practical way to calculate text length and time variability and measure reliable change when assessing clients. *Topics in Language Disorders, 40*(2), 166-181.

Stokes, S. F., Wong, A. M., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research*, *49*(2), 219-236.

Sung, Y. T., Cha, J. H., Tu, J. Y., Wu, M. D., & Lin, W. C. (2016). Investigating the processing of relative clauses in Mandarin Chinese: evidence from eye-movement data. *Journal of psycholinguistic research*, *45*(5), 1089-1113.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological science in the public interest*, *1*(1), 1-26.

Tardif, T., & Fletcher, P. (2008). Chinese Communicative Development Inventories: user's guide and manual.

Thal, D. J., O'Hanlon, L., Clemmons, M., & Fralin, L. (1999). Validity of a parent report measure of vocabulary and syntax for preschool children with language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(2), 482-496.

Thompson, S. A., & Tao, H. (2010). Conversation, grammar, and fixedness: adjectives in Mandarin revisited. *Chinese Language and Discourse*, *1*(1), 3-30.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1284-1294.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of speech, language, and hearing research*, *40*(6), 1245-1260.

Trauner, D., Wulfeck, B., Tallal, P., & Hesselink, J. (2000). Neurological and MRI profiles of children with developmental language impairment. *Developmental Medicine & Child Neurology*, *42*(7), 470-475.

Tsimpli, I. M., Andreou, M., & Peristeri, E. (2020). The multilingual assessment instrument for narratives: Greek. *ZAS Papers in Linguistics*, *64*, 101-106.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323-352.

US Preventive Services Task Force. (2006). Screening for speech and language delay in preschool children: recommendation statement. *Pediatrics*, *117*(2), 497-501.

Van der Lely, H. K., & Battell, J. (2003). Wh-movement in children with grammatical SLI: A test of the RDDR hypothesis. *Language*, 153-181.

Vân Hoàng, T., Schelstraete, M. A., Trần, Q. D., & Bragard, A. (2014). La 46epetition de phrases en vietnamien–un marqueur des troubles du langage oral et des troubles du comportement Sentence repetition in Vietnamese-a marker of oral language and behavioral difficulties. *Canadian Journal of Speech-Language Pathology and Audiology*, *37*(4), 280-297.

Wang, W. S., & Sun, C. (2015). *The Oxford handbook of Chinese linguistics*. Oxford University Press.

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Language, and Hearing Research*, *38*(6), 1349-1355.

Wechsler, D. (1989). Wechsler preschool and primary intelligence scales for children, revised. *New York, NY, The Psychological Corporation*.

Wiedenhof, J. (2015). *A grammar of Mandarin*. John Benjamins Publishing Company.

Wiig, E., Semel, E., & Seccord, W. (2013). Clinical Evaluation of Language Fundamentals-Revised. *NY: Merrill Divison of Macmillan*.

Woon, C. P., Yap, N. T., Lim, H. W., & Wong, B. E. (2014). Measuring Grammatical Development in Bilingual Mandarin-English Speaking Children with a Sentence Repetition Task. *Journal of Education and Learning*, *3*(3), 144-157.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32-35.

Zeng, T., Zhu, T., Li, X., & Zhu, R. (2018). Passive Structure Features in Mandarin-Speaking Children with Specific Language Impairment: Optional Movement. *Journal of Language, Linguistics, and Literature, 4(1)*, 8-18.

Tables and Figures

Table 1. Correlations between SR scoring systems and validation measures

	MLU	VOCD	Number of predicates	Structure Composite
Binary	.435*	.356*	.380*	.594*
Core element	.246	.342*	.148	.397*
Error method – syllable	.432*	.402*	.389*	.587*
Error method – word	.419*	.400*	.392*	.599*
Percent of correct syllable	.383*	.441*	.379*	.559*

Note: p level is corrected using the Bonferroni correction, and the corrected p value is .003

* $p<.003$

Table 2. The demographic characteristics and standardized test scores of the TD and DLD groups

Measure	TD (N=16)		DLD (N=16)		t/W	p value	Cohen's d
	Mean (SD)	Range	Mean (SD)	Range			
Age	60.8 (6.5)	50-71	60.3 (6.8)	50-70	*t*=.10	.92	.05
Mandarin Exposure	3.8 (1.0)	1-5	3.9 (1.1)	1-5	*W*=254	.68	.06
Maternal Education	4.2 (.8)	3-5	3.8 (.9)	3-5	*W*=232	.19	.52
Performance IQ	120.9 (11.7)	99-139	101.9 (11.6)	87-124	*t*=4.5	<.001	1.63
DREAM total score	109.7 (8.2)	94-123	84.8 (7.6)	72-103	*t*=8.9	<.001	3.14

Note: Value t is reported for t-tests when the two variables under comparison were normally distributed; W is reported for Mann Whitney U tests when one or more of the variables under comparison were not normally distributed. Age is reported in months. Mandarin exposure is reported on a scale from 1 to 5: 1 means <19%; 2 means 20%-39%; 3 means 40%-59%, 4 means 60%-79%, and 5 means 80%-100%. One TD child and one child with DLD received a Mandarin exposure score of 1. The TD child had exposure to Shanghai dialect at home and attended English classes four to five hours each week. The child with DLD had exposure to both Shanghai and Henan dialects at home and attended English classes for half a year. Performance IQ is measured by WPPSI-R and is reported as standard scores. Maternal education is reported on a scale from 1 to 5 through a parent questionnaire: 1 means middle school or lower, 2 means high school, 3 means some college, 4 means bachelor's degree, and 5 means master's degree. The total scores on DREAM are reported as standard scores.

Figure captions:

Figure 1. Children's SR scores using the five different scoring methods

Figure 2. Children's MSRT accuracy score by group and scoring method

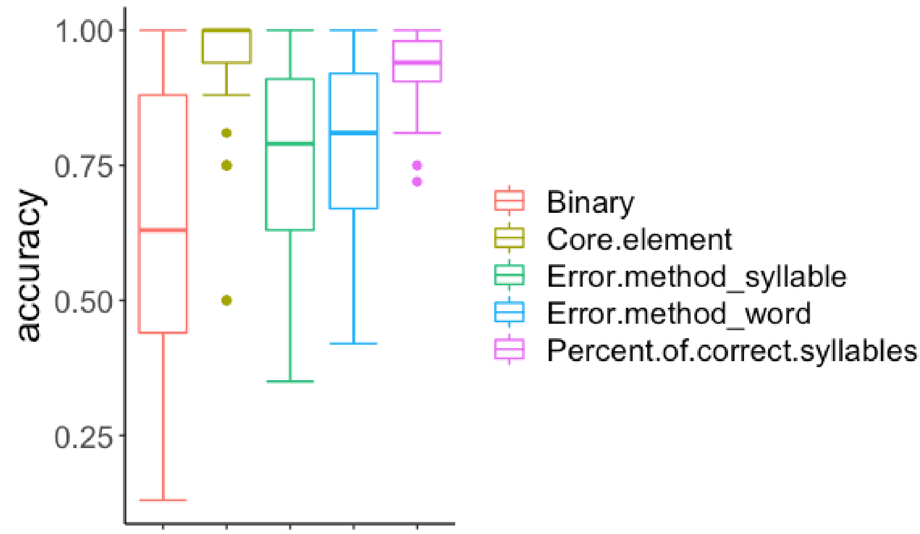
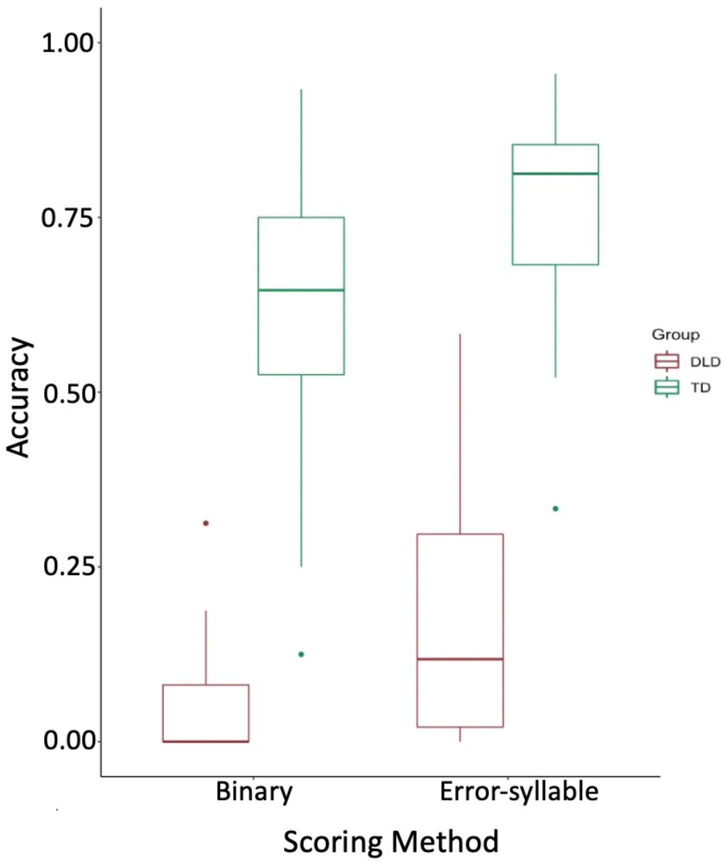


Figure 1. Children's SR scores using the five different scoring methods


1.00
0.75
0.50
0.25
0.00
Accuracy
Binary
Error-syllable
Scoring Method
Group
DLD
TD

Appendix A. Sentences in the MSRT

Classifiers are bolded and relative clauses are underlined.

PASS=passive marker; SFP=sentence final particle; ProM = progressive marker; PerM = perfective marker; LP=linking particle; RP=resultative particle。

Passive Sentences:

1. 那 **只** 白色 的 小狗 被 妈妈 抱 走 了。
   That CL white LP dog PASS mom carry away (RP) SFP
   (That white dog was carried away by mom.)
2. 那 **件** 破旧 的 毛衣 被 姐姐 扔 掉 了。
   That CL old LP sweater PASS sister throw away (RP) SFP
   (That old sweater was thrown away by sister.)
3. 男孩 被 那 个 穿 裙子 的 女孩 绊 倒 了。
   Boy PASS that CL wear dress LP girl tripped fell (RP) SFP
   (The boy was tripped by that girl who wore a dress.)
4. 大灰狼 被 那 **只** 聪明 的 小羊 打 败 了。
   Wolf PASS that CL smart LP sheep defeated lost (RP) SFP
   (The wolf was defeated by that smart sheep.)
5. 蛋糕 被 那 个 戴 眼镜 的 女孩 吃 光 了。
   Cake PASS that CL wear glasses LP girl eaten up (RP) SFP
   (The cake was eaten by the girl who wore glasses.)
6. 那 **辆** 红色 的 自行车 被 小偷 偷 走 了。
   That CL red LP bicycle PASS thief stole away (RP) SFP
   (That red bicycle was stolen by the thief.)
7. 那 个 长 头发 的 女孩 被 一 **只** 猫 抓 伤 了。
   That CL long hair LP girl PASS one CL cat scratched hurt (RP) SFP
   (That girl with long hair was scratched by a cat.)
8. 棒棒糖 被 那 个 高个子 的 男孩 抢 走 了。
   Lollipop PASS that CL tall LP boy robbed away (RP) SFP
   (The lollipop was robbed by that tall boy.)

Aspect marker sentences:

1. 那 个 戴 帽子 的 男孩 在 骑 自行车。
   That CL wear hat LP boy ProM ride bicyle
   That boy who wears a hat is riding a bicycle.
2. 哥哥 在 组装 那 **辆** 绿色 的 玩具 车。
   Older brother ProM assemble that CL green LP toy car
   Older brother is assembling that green toy car.
3. 那 **群** 中班 的 小朋友 在 高兴 地 荡秋千。
   Those CL kindergarten LP children ProM happily LP play on a swing
   Those kindergarten children are playing on a swing happily.
4. 老师 在 严厉 地 批评 那 个 淘气 的 男孩。
   Teacher ProM harshly LP criticize that CL naughty LP boy

The teacher is harshly criticizing that naughty boy.

LE:

1. 小 老鼠 吃 了 一 **块** 美味 的 巧克力。
   Little mouse eat PerM one CL delicious LP chocolate
   The little mouse has eaten a piece of delicious chocolate.
2. 那 个 背 书包 的 男孩 掉 了 一 **本** 书。
   That CL carry bag LP boy drop PerM one CL book
   That boy who carries a bag has dropped a book.
3. 那 个 可爱 的 女孩 大声 地 唱 了 一 **首** 歌。
   That CL cute LP girl loudly LP sing PerM one CL song
   That cute girl has sung a song loudly.
4. 妈妈 认真 地 洗 了 一 **件** 漂亮 的 衣服。
   Mother seriously LP wash PerM one CL beautiful LP clothes
   Mother has seriously washed a beautiful clothes.

A. Scatter plots between the SR scores and the narrative measures in study 1

Figure 1. Scatter Plots of the Error Method Word Scoring and Narrative Measures

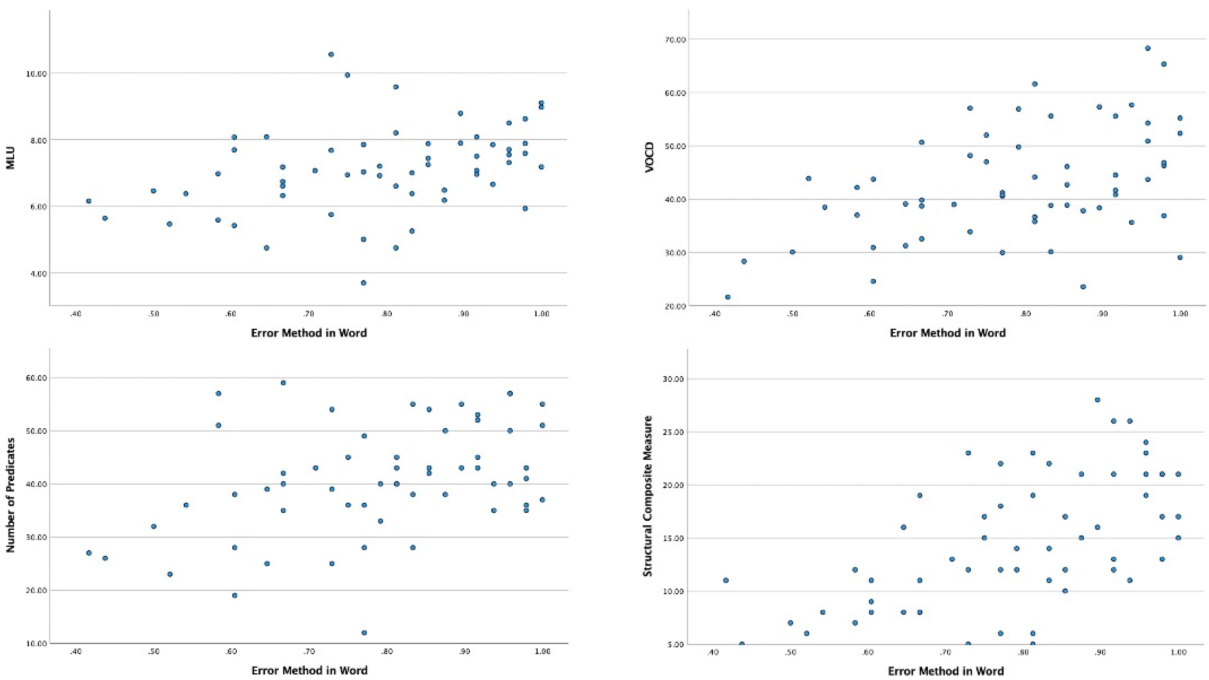


Figure 2. Scatter Plots of the Error Method Syllable Scoring and Narrative Measures

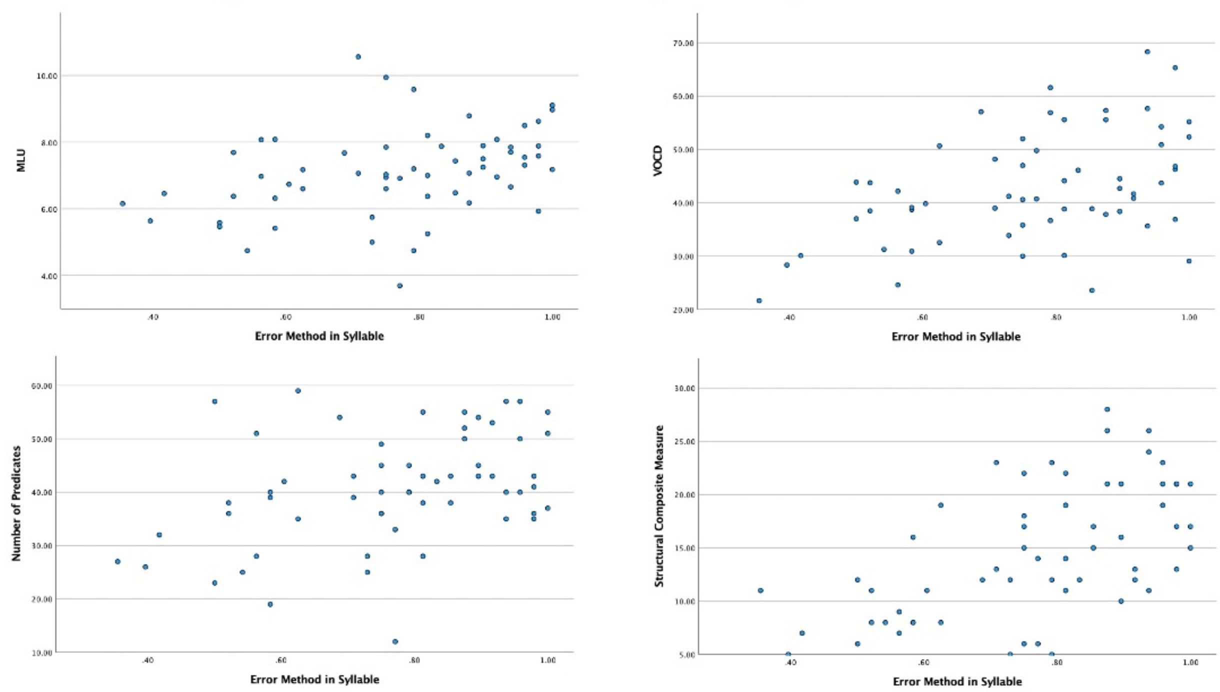


Figure 3. Scatter Plots of the Binary Scoring and Narrative Measures

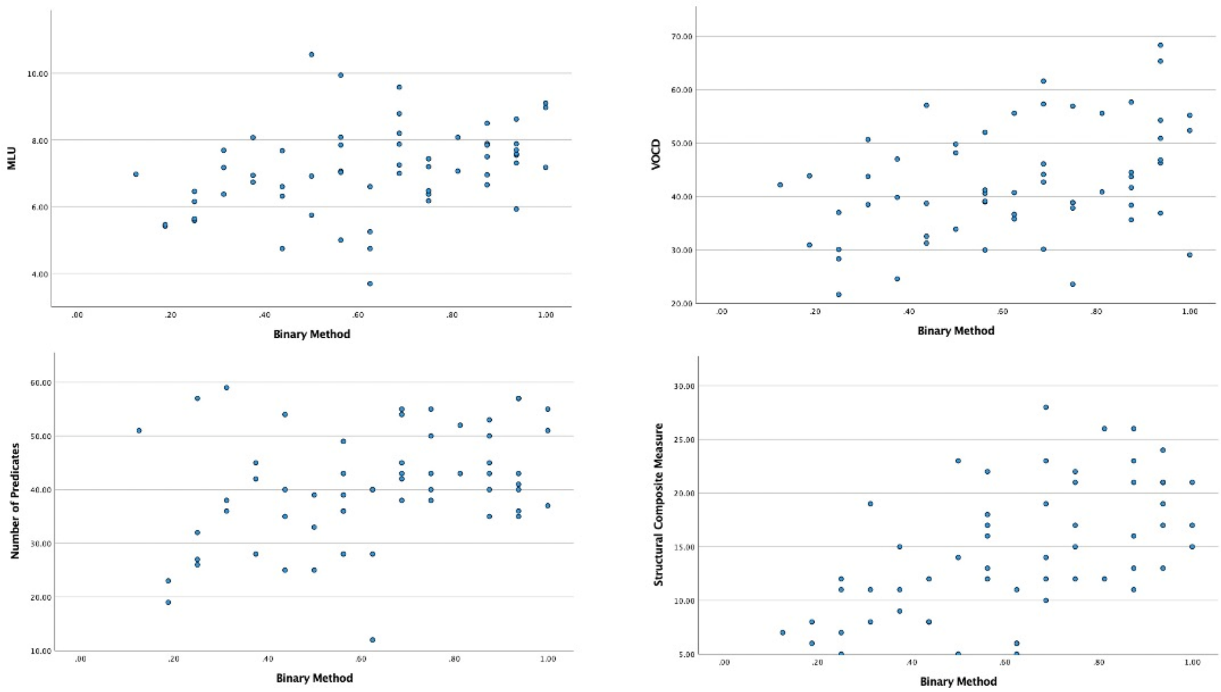


Figure 4. Scatter Plots of the Core Element Scoring and Narrative Measures

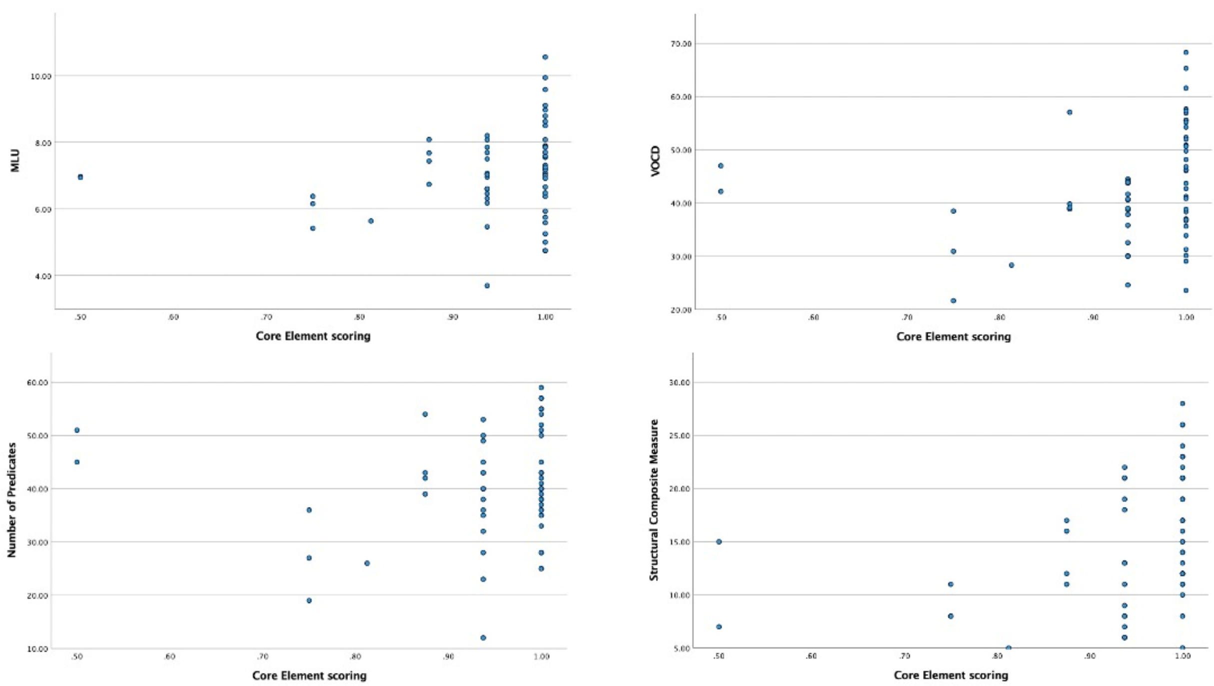
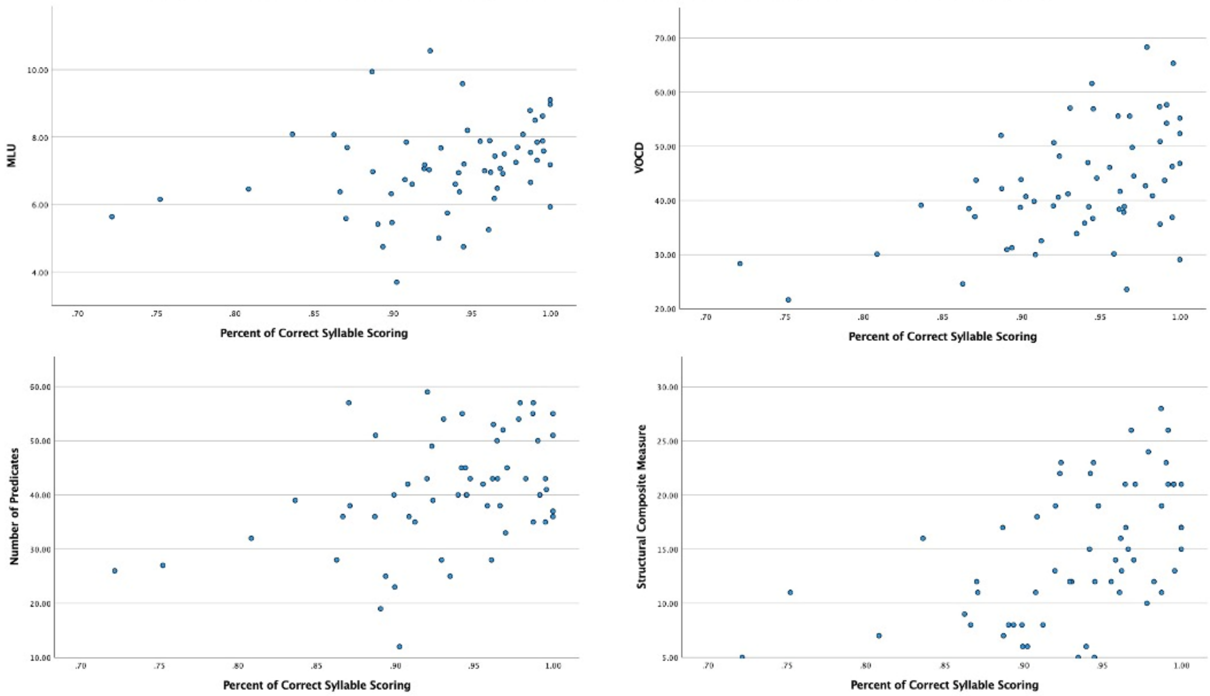


Figure 5. Scatter Plots of the Percent of Correct Syllable Scoring and Narrative Measures



B. Examples of a sample response and its scoring procedures

Target sentence:
小 老鼠 吃 了 一 块 美味的 巧克力。
xiao3 lao3shu3 chi1 le yi1 kuai4 mei3we4de qiao3ke1li4
Little mouse eat PerM one CL delicious chocolate
The little mouse has eaten a piece of delicious chocolate.

Child's response:
小 老鼠 吃 了 一 块 ***好吃***的 巧克力。
xiao3 lao3shu3 chi1 le yi1 kuai4 ***hao3chi1***de qiao3ke1li4
Little mouse eat PerM one CL ***tasty*** chocolate
The little mouse has eaten a piece of ***tasty*** chocolate.

Error method – word: The child substituted the word "美味" with "好吃", which was considered as one word error. A score of **2** (1 error) was given for this response using the error method in word system.

Error method – syllable: The child substituted two syllables "美味" with "好吃", which was considered as two syllable errors. A score of **1** (two to three errors) was given for this response using the error method in syllable system.

Binary method: As this response was not a completely accurate repetition of the target, a score of **0** was given using the binary method.

Core element scoring: The core element in the target sentence is the aspect marker structure (verb + aspect marker + noun). The target core element is present in the child's response, so a score of **1** was given using the core element scoring.

Percent of correct syllable scoring: The target sentence has a total of 13 syllables and 11 of them were repeated correctly in the response. The percent of correct syllable scoring thus gives a score of 11/13, which is **0.85**.

C. Descriptive Data of Children's Scores on MSRT and Narrative Measures in Study 1

Measure	Range	Mean	SD
Error method – word	0.42-1.00	0.79	0.15
Error method – syllable	0.35-1.00	0.77	0.17
Binary method	0.13-1.00	0.63	0.25
Core element method	0.50-1.00	0.94	0.11
Percent of correct syllable method	0.72-1.00	0.93	0.06
MLU-w	3.69-10.56	7.11	1.30
VOCD	21.65-68.32	42.75	10.39
Number of predicates	12-59	40.69	10.40
Structural composite	5-28	14.63	6.14

D. Examples of the narrative measures

Structure type	Subtype	Example
Predicate	Verb	然后 小 鸟 妈妈 **爬** 到 树上。 Then little bird mom **climbed** onto the tree Then, the little bird mom climbed onto the tree.
	Predicate adjective	小猫 真 **生气**。 The cat so **angry** The cat is so angry.
Aspect Marker	Progressive	小狗 **在** 捉 老鼠。 The dog ProM catch mouse The dog is catching the mouse.
	Perfective	然后 小 男孩 拣回 **了** 自己的 球。 Then little boy got back PerM his ball Then the little boy got his ball back.
Classifier	-	小猫 吃 了 一 **条** 鱼。 The cat ate PerM one CL fish The cat at a fish.
Passive	-	它的 腿 **被** 狐狸 吃 了。 Its leg PASS fox eat SFP Its leg was eaten by the fox.
Relative Clause	-	看到 **掉 到 河里 的 球** 了。 Saw dropped RP river LP ball SFP Saw the ball that dropped in the water.

Note: ProM=progressive marker; PerM=perfective marker; CL=classifier; PASS=passive marker; SFP=sentence final particle; RP=resultative particle; LP=linking particle.